DataSet
Santander

Processo Seletivo - Daniel de J. Siqueira

Overview

1. Objetivos
2. Estratégia
3. Resultados
4. Aplicação

serasa experian™

Objetivos

	target	var_0	var_1	…	var_197	var_198	var_199
0	0	8.9255	-6.7863	…	8.5635	12.7803	-1.0914
1	0	11.5006	-4.1473	…	8.7889	18356	1.9518
…	…	…	…	…	…	…	…
199998	0	9.7148	-8.6098	…	10.0342	15.5289	-13.9001
199999	0	10.8762	-5.7105	…	8.1857	12.1284	0.1385

serasa experian™

Objetivos

- Redirecionamento de tendências
- Targeted Marketing
- Valores do Santander

serasa experian™

Estratégia

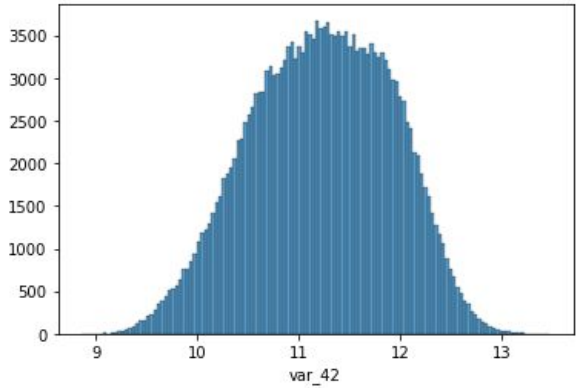
Vamos usar Machine Learning!!

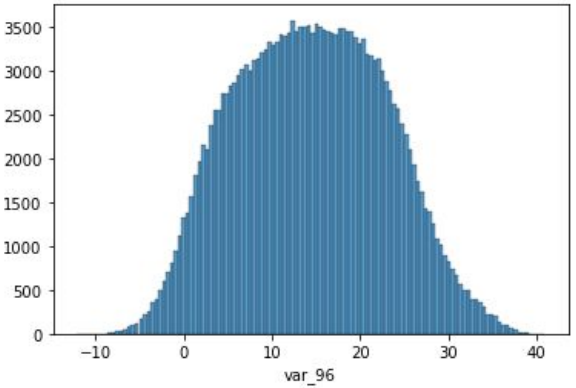
Primeiro, deveríamos fazer um tratamento dos dados, mas o dataset já estava limpo. 🙂👍

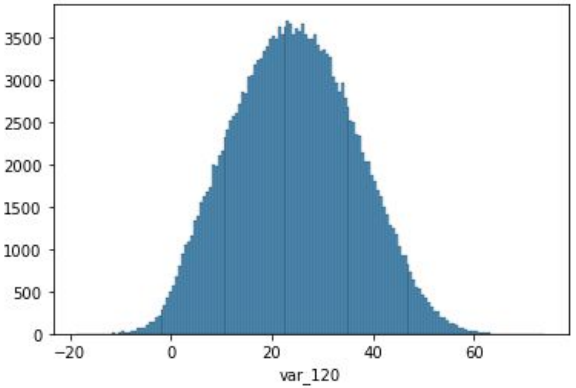
serasa experian™

Estratégia

Em seguida, procuramos padrões nas variáveis.







Maior Correlação: Corr(var_26,var_139) = 0.0098.

serasa experian™

Estratégia

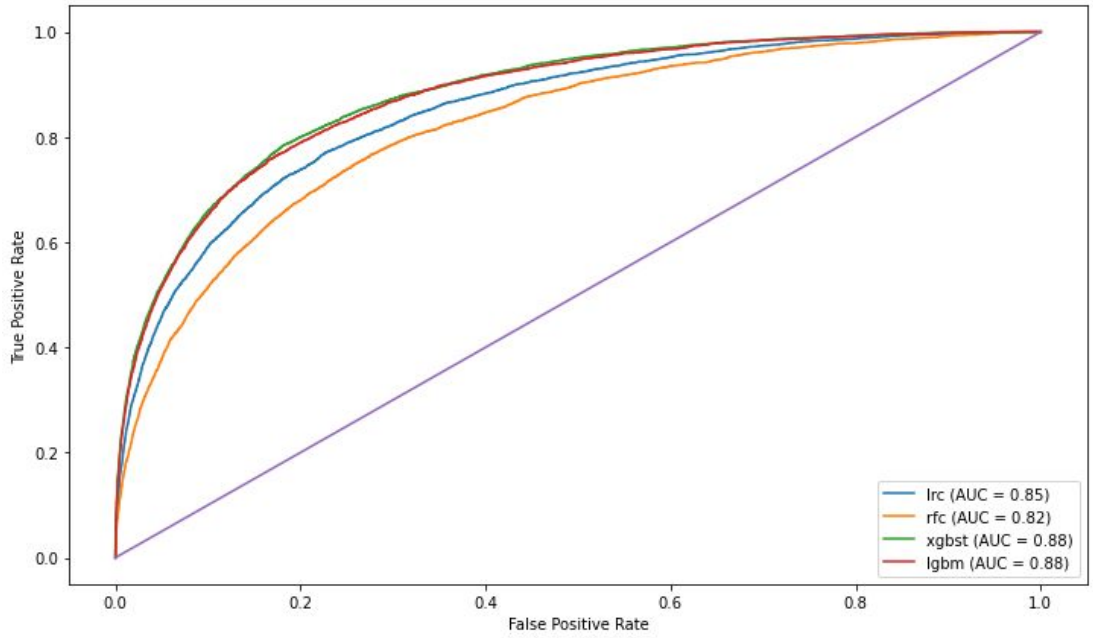
Modelo: GridSearchCV + Undersampling +

- **LogisticRegression;**
- **RandomForest;**
- **XGBoost;**
- **LightGBM.**

serasa experian

Resultados

Métrica escolhida: ROC AUC

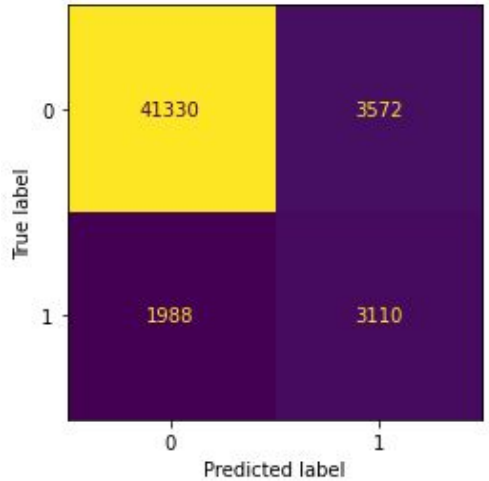


serasa experian

Aplicação

Como aplicar o modelo escolhido, na prática?

Para o XGBoost com t = 0.5, a matriz de confusão é:

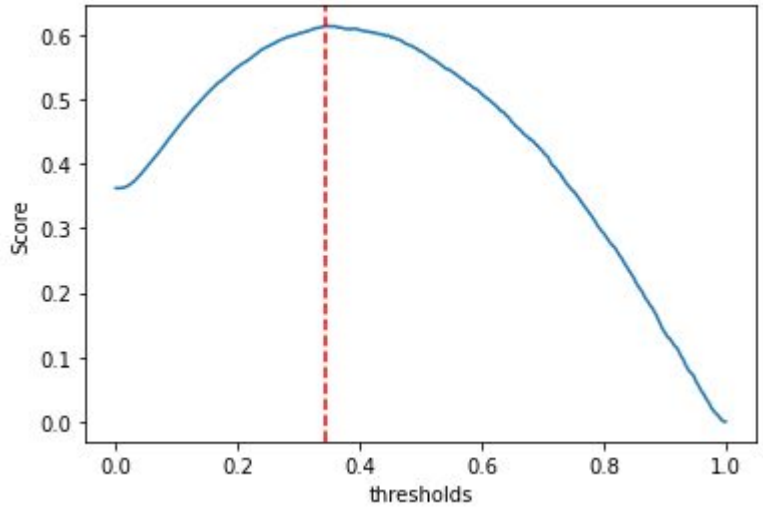


serasa experian

Aplicação

Uma possível escolha de custo:

F2 score = Média Harmônica entre Recall (com peso 2) e Precision.



- t = 0.34
- F2 score = 0.61

Antes (t = 0.5):

- F2 score = 0.57

serasa experian

Conclusão

- Desempenho final
- Possíveis melhorias
- Caminhos não explorados

serasa experian